Support Vector Regression (1)

Manel Martínez-Ramón

ECE, UNM

October 2018

- The SVM approach is valid for this task.
- The criterion is the same:
  - Find a set of constraints that define a set of loss or slack variables to define an empirical risk.
  - Minimize the empirical risk and the complexity in the same functional.

Linear SVR model

- The regression model is defined as

$$y_n = \mathbf{w}^\top \mathbf{x}_n + b + e_n$$

  where $y_n \in \mathbb{R}$ is the set of regressors (desired outputs), and e_n is the regression error for sample $\mathbf{x}_n$
- The error is also continuous variable in $\mathbb{R}$, so the definition of positive slack variables must be defined as a function of the error norm

Constraints

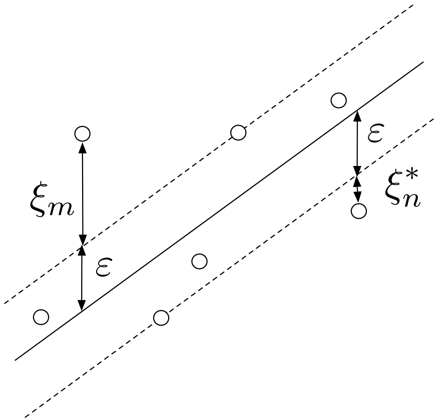
- In order to define positive slack variables, we consider the positive and negative error cases (Smola, 2003):

$$
\begin{aligned}
y_n - \mathbf{w}^\top \mathbf{x}_n - b &\leq \varepsilon + \xi_n \\
-y_n + \mathbf{w}^\top \mathbf{x}_n + b &\leq \varepsilon + \xi_n^* \\
\xi_n, \xi_n^* &\geq 0
\end{aligned}
$$

- Interpretation:
  - We define an ε margin or ε tube as an error tolerance $\pm\varepsilon$.
  - If the error is less than $|\varepsilon|$, the slacks are dropped to zero.
  - Otherwise, the slacks are positive.
- We place as many samples as possible *inside* the margin while minimizing the loss on the samples *outside* the margin.

Constraints

- The samples inside the margin have zero slack variables.



- The original development of the SVR minimizes the sum of the slack variables, which is a linear risk, plus the complexity.

$$\text{Minimize } ||\mathbf{w}||^2 + C \sum_{n=1}^{N} (\xi_i + \xi_i^*)$$

$$\text{subject to } \begin{cases} y_n - \mathbf{w}^\top \mathbf{x}_n - b - \varepsilon - \xi_n \leq 0 \\ -y_n + \mathbf{w}^\top \mathbf{x}_n + b - \varepsilon - \xi_n^* \leq 0 \\ \xi_n, \xi_n^* \geq 0 \end{cases}$$

The SVR Lagrangian

- The Lagrange optimization involves four sets of positive Lagrange multiplers: α_n, α_n^* and μ_n, μ_n^*.
- We must *maximize* the two first constraints, minimize the other two. The Lagrangian is

$$
\begin{aligned}
L_L = ||\mathbf{w}||^2 + C\sum_{n=1}^{N}(\xi_i + \xi_i^*) \\
-\sum_{n=1}^{N}\alpha_i(-y_n + \mathbf{w}^\top\mathbf{x}_n + b + \varepsilon + \xi_n) \\
-\sum_{n=1}^{N}\alpha_i^*(y_n - \mathbf{w}^\top\mathbf{x}_n - b + \varepsilon + \xi_n^*) \\
-\sum_{n=1}^{N}(\mu_n\xi_n + \mu_n^*\xi_n^*)
\end{aligned}
$$

- A procedure similar to the one used for the SVC, that takes into account the KKT complementary conditions for the product constraint-Lagrange multipliers, leads to the minimization of

$$L_d = -\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top \mathbf{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top \mathbf{y} - \varepsilon \mathbf{1}^\top(\boldsymbol{\alpha} + \boldsymbol{\alpha}*)$$

  with $0 \leq \alpha_n, \alpha_n^* \leq C$
- This is a quadratic form that has the property of existence and uniqueness of a solution.
- The optimization of this functional is almost exactly equal to the one needed for the SVC.

SVR dual

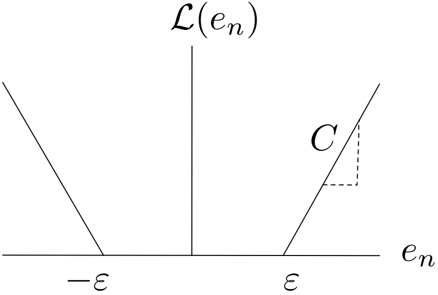
- The set of parameters is a function of the data

$$\mathbf{w} = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*) \mathbf{x}_n$$

- The bias is obtained the same way as the one of the SVC, using the KKT conditions.
- The samples inside the margin have null dual variables α_n. The ones on the margin or outside have positive values in their dual variables α_n.

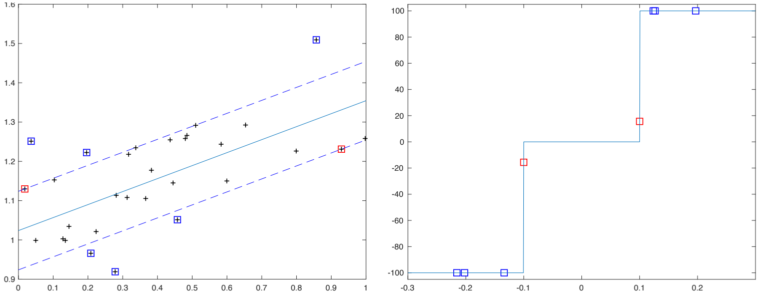
The Vapnik or ε-insensitive cost function

- The implicit cost or loss function over the errors is linear outside the margin.



- It can be shown that the values of the dual variables are the derivative of the cost function.

Example of a linear SVR



Numerical regularization

- This numerical regularization has an interpretation as a modification of the cost function. The new dual to be maximized is

$$L_d = -\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top (\mathbf{K} + \gamma \mathbf{I})(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top \mathbf{y} - \varepsilon \mathbf{1}^\top (\boldsymbol{\alpha} + \boldsymbol{\alpha}^*)$$

- This is not exactly the result produced by the previous Lagrangian.
- Nevertheless, the use of a modified cost function leads exactly to this result and has an interpretation in the way in which the errors are actually processed (Rojo-Álvarez et al., 2004).

- The implicit cost function applied in practice is

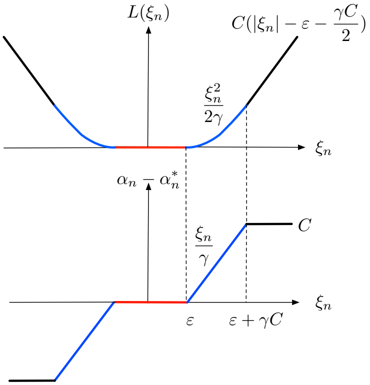
$$L(\xi_n) = \begin{cases} C(|\xi_n| - \varepsilon - \frac{\gamma C}{2}) & |\xi_n| > \varepsilon + \gamma C \\ \frac{1}{2\gamma}(|\xi_n| - \varepsilon)^2 & 0 < |\xi_n| < \varepsilon + \gamma C \\ 0 & |\xi_n| < \varepsilon \end{cases}$$

Whose derivative is

$$\alpha_n(\xi_n) - \alpha_n^*(\xi_n) = \begin{cases} C & |\xi_n| > \varepsilon + \gamma C \\ \frac{1}{\gamma}|\xi_n| & 0 < |\xi_n| < \varepsilon + \gamma C \\ 0 & |\xi_n| < \varepsilon \end{cases}$$

The ε-Huber cost function

- The function is similar to the one of the SVC and is a combination of the ε-insensitive cost function plus the robust Huber cost function.
- The data inside the quadratic part must be the one likely to be Gaussian. The linear part will process the outliers.



Outcomes of this lesson

- Application of the SVM criterion to regression: reinterpretation of the margin.
- Practical optimization of the SVR.
- Properties of the dual variables in relation to their corresponding support vectors.
- Interpretation (again) of the matrix regularization (ε-Huber cost function).